



SUBSTITUTE SPECIFICATION CLEAN VERSION

METHODS AND SYSTEMS FOR ADJUSTING A SCORING MEASURE BASED ON QUERY BREADTH

FIELD OF THE INVENTION

[0001] The present invention relates generally to methods and systems for information retrieval. The present invention relates particularly to methods and systems for adjusting a scoring measure associated with a search result based on the breadth of a previously-executed search query associated with the search result.

BACKGROUND

[0002] A conventional network search engine, such as the Google™ search engine, returns a result set in response to a search query submitted by a user. The search engine performs the search based on a conventional search method. For example, one known method, described in an article entitled “The Anatomy of a Large-Scale Hypertextual Search Engine,” by Sergey Brin and Lawrence Page, assigns a degree of importance to a document, such as a web page, based on the link structure of the web page. The search engine ranks or sorts the individual articles or documents in the result set based on a variety of measures. For example, the search engine often ranks the results based on a popularity measure. The search engine generally places the most popular results at the beginning of the result set.

[0003] The popularity measure may comprise one or more individual popularity measures. For example, a search engine may utilize the number of times a particular document has been shown to users, i.e., impression count, as a measure of popularity. A conventional search engine may also use a click count or click-through

ratio as a measure of popularity. While these measures provide valuable information about each result, the measures can be insufficient, depending on a variety of factors.

[0004] A search engine often retrieves a large number of documents for a broad query. For example, if a user enters a one or two-term query, such as “digital camera,” the search engine is likely to return millions of results. Also, many different users may submit this broad query initially when searching about material related to digital cameras. Accordingly, the documents returned by these broad queries are often over-represented in the popularity counts, and the popularity count for each one of these results is artificially high because of the number of broad queries submitted. Also, documents returned in response to broad queries are often more abstract than results returned for more specific queries. The more abstract documents are then over-represented in the popularity counts, whether based on clicks or based on impressions.

[0005] The resulting over-representation of documents due to broad queries tends to skew data collected about the users’ behavior. When a user views a result set from a very broad query, the user will likely see only a small fraction of the entire result set. Therefore, it is difficult to, for example, determine the popularity of documents in the result set based on the users’ response to documents resulting from a broad query.

SUMMARY

[0006] Embodiments of the present invention provide methods and systems for adjusting a scoring measure for a search result based at least in part on the breadth of a previously-executed search query associated with the result. In one embodiment, a

search engine determines a popularity measure for a search result and adjusts the popularity measure based at least in part on a query breadth measure of a previously-executed search query associated with the search result. The search engine may use a variety of query breadth measures. For example, in one embodiment, the query breadth measure comprises a quantity of results returned by the search query.

[0007] These exemplary embodiments are mentioned not to limit or define the invention, but to provide examples of embodiments of the invention to aid understanding thereof. Exemplary embodiments are discussed in the Detailed Description, and further description of the invention is provided there. Advantages offered by the various embodiments of the present invention may be further understood by examining this specification.

BRIEF DESCRIPTION OF THE FIGURES

[0008] These and other features, aspects, and advantages of the present invention are better understood when the following Detailed Description is read with reference to the accompanying drawings, wherein:

Figure 1 is a block diagram illustrating an exemplary environment in which one embodiment of the present invention may operate;

Figure 2 is a flowchart illustrating a process for associating a popularity measure with a search result in one embodiment of the present invention;

Figure 3 is a flowchart illustrating a process for associating a query breadth measure with a search query in one embodiment of the present invention; and

Figure 4 is a flowchart illustrating a method for adjusting the popularity measure of a result in one embodiment of the present invention.

DETAILED DESCRIPTION

[0009] Embodiments of the present invention comprise methods and systems for adjusting a scoring measure for a search result based at least in part on the breadth of a previously-executed search query associated with the search result. In one embodiment, a search engine deweights a popularity measure for a result when the value of the popularity measure has been increased as a result of the submission of one or more broad queries. The breadth of the query may be calculated in various ways.

[0010] Referring now to the drawings in which like numerals indicate like elements throughout the several figures, Figure 1 is a block diagram illustrating an exemplary environment for implementation of one embodiment of the present invention. The system 100 shown in Figure 1 includes multiple client devices 102a-n in communication with a server device 104 over a network 106. The network 106 shown includes the Internet. In other embodiments, other wired and wireless networks, such as an intranet may be used. Moreover, methods according to the present invention may operate within a single computer.

[0011] The client devices 102a-n shown each includes a computer-readable medium, such as a random access memory (RAM) 108 coupled to a processor 110. The processor 110 executes computer-executable program instructions stored in memory 108. Such processors may include a microprocessor, an ASIC, and state machines. Such processors include, or may be in communication with, media, for example computer-readable media, which stores instructions that, when executed by the processor, cause the processor to perform the steps described herein.

Embodiments of computer-readable media include, but are not limited to, an electronic, optical, magnetic, or other storage or transmission device capable of providing a processor, such as the processor 110 of client 102a, with computer-readable instructions. Other examples of suitable media include, but are not limited to, a floppy disk, CD-ROM, DVD, magnetic disk, memory chip, ROM, RAM, an ASIC, a configured processor, all optical media, all magnetic tape or other magnetic media, or any other medium from which a computer processor can read instructions. Also, various other forms of computer-readable media may transmit or carry instructions to a computer, including a router, private or public network, or other transmission device or channel, both wired and wireless. The instructions may comprise code from any computer-programming language, including, for example, C, C++, C#, Visual Basic, Java, Python, Perl, and JavaScript.

[0012] Client devices 102a-n may also include a number of external or internal devices such as a mouse, a CD-ROM, DVD, a keyboard, a display, or other input or output devices. Examples of client devices 102a-n are personal computers, digital assistants, personal digital assistants, cellular phones, mobile phones, smart phones, pagers, digital tablets, laptop computers, Internet appliances, and other processor-based devices. In general, a client device 102a may be any type of processor-based platform that is connected to a network 106 and that interacts with one or more application programs. Client devices 102a-n may operate on any operating system capable of supporting a browser or browser-enabled application, such as Microsoft® Windows® or Linux. The client devices 102a-n shown include, for example, personal computers executing a browser application program such as Internet

Explorer™ from Microsoft Corporation, Netscape Navigator™ from Netscape

Communications Corporation, and Safari™ from Apple Computer, Inc.

[0013] Through the client devices 102a-n, users 112a-n can communicate over the network 106 with each other and with other systems and devices coupled to the network 106. As shown in Figure 1, a server device 104 is also coupled to the network 106. In the embodiment shown, a user 112a-n generates a search query 114 at a client device 102a. The client device 102a transmits the query 114 to the server device 104 via the network 106. For example, a user 112a types a textual search query into a query field of a web page of a search engine interface displayed on the client device 102a, which is then transmitted via the network 106 to the server device 104. In the embodiment shown, a user 112a inputs a search query 114 at a client device 102a, which transmits a search query signal 130 associated with the search query 114 to the server device 104. The search query 114 may be transmitted directly to the server device 104 as shown. In another embodiment, the query signal 130 is instead sent to a proxy server (not shown), which then transmits the query signal 130 to server device 104. Other configurations are possible.

[0014] The server device 104 shown includes a server executing a search engine application program, such as the Google™ search engine. Similar to the client devices 102a-n, the server device 104 shown includes a processor 116 coupled to a computer-readable memory 118. Server device 104, depicted as a single computer system, may be implemented as a network or cluster of computer processors. Examples of a server device 104 are servers, mainframe computers, networked computers, a processor-based device, and similar types of systems and devices.

Client processor 110 and the server processor 116 can be any of a number of computer processors, such as processors from Intel Corporation of Santa Clara, California and Motorola Corporation of Schaumburg, Illinois.

[0015] Memory 118 contains the search engine application program, also known as a search engine 120. The search engine 120 locates relevant information in response to a search query 114 from a user 112a-n. The search engine 120 then provides the result set 134 to the client 102a via the network 106.

[0016] In the embodiment shown, the server device 104, or related device, has previously performed a crawl of the network 106 to locate articles, such as web pages, stored at other devices or systems connected to the network 106, and indexed the articles in memory 118 or on another data storage device. Articles include, for example, web pages of various formats, such as HTML, XML, XHTML, Portable Document Format (PDF) files, and word processor, database, and application program document files, audio, video, or any other documents or information of any type whatsoever made available on a network (such as the Internet), a personal computer, or other computing or storage means. The embodiments described herein are described generally in relation to HTML files or documents, but embodiments may operate on any type of article, including any type of image.

[0017] The search engine 120 includes a document locator 122, a ranking processor 124, and a query breadth analyzer 126. In the embodiment shown, each comprises computer code residing in the memory 118. The document locator 122 identifies a set of documents that are responsive to the search query 114 from a user 112n by, for example, accessing an index of documents, indexed in accordance with

potential search queries or search terms. The ranking processor 124 ranks or scores the search result 134 including the located set of web pages or documents based upon relevance to the search query 114 or another criteria, such as a popularity measure. The query breadth analyzer 126 determines or otherwise measures the breadth of the query associated with the query signal 130. Note that other functions and characteristics of the document locator 122, ranking processor 124, and query breadth analyzer 126 are further described below.

[0018] Server device 104 also provides access to other storage elements, such as a popularity database 128 and a query breadth database 129. The server device 104 can access other similar types of data storage devices. The popularity database 128 stores measures of the popularity of a document. The popularity database 128 may contain either or both query-dependent and query-independent popularity measures. For example, the popularity database 128 may store the number of times a particular document has been shown to users - a query-independent measure. Alternatively or additionally, the popularity database 128 may store the number of times that the article has been shown to a user in response to a particular keyword - a query-dependent measure. The ranking processor 124 utilizes this information in performing rankings of pages in the result set 134. When the search engine 120 generates a result set 134, the search engine 120 causes a record to be added or modified in the popularity database 128 to indicate that a particular result was shown to a user 112a in a result set 134.

[0019] Other measures of popularity may be stored in the popularity database 128, including, for example, click-through data. Click-through data is generally an

indicator of quality in a search result. Quality signals or click-through data can include, but is not limited to, whether a particular URL or document is clicked by a particular user; how often a URL, document, or web page is clicked by one or more users; how often a particular user clicks on specific documents or web pages; and the ratio of how often a URL, document, or web page is clicked by one or more users to the number of times the URL, document or web page is shown to one or more users (known also as the click-through ratio). A popularity database 128 or similar data storage devices can store other types of quality signals similar to click-through data, such as any quantitative measure of user behavior.

[0020] Other data related to documents located in a search result 134 that can be stored in a popularity database 128 or other data storage device can include, but is not limited to, how often a particular URL, document, or web page is shown in response to a search query 114; how many times a particular search query 114 is asked by users 112; the age or time a particular document has been posted on a network 106, and the identity of a source of a particular document on a network 106.

[0021] The query breadth database 129 stores measures of query breadth. Various measures of query breadth may be stored. For example, one measure of query breadth is the number of results matching a query. The higher the number of results, the more likely the query is a broad one.

[0022] Another measure of query breadth is the rate at which the information retrieval (IR) score drops off from a first result until a second result, e.g., the first result until the nth result. The terms first and second are used here merely to differentiate one item from another item. The terms first and second are not used to

indicate first or second in time, or first or second in a list, or other order, unless explicitly noted. The IR score provides a measure of the relevance of a document for a query. If many documents are returned in response to a query, many of them are likely relevant to the query and will have a correspondingly high IR score.

Accordingly, the IR score will drop slowly from one result to the next result in the result set. The slower the rate of drop off, the more likely the query is a broad query.

For example, if a query returns one million results, the drop off in IR score from the first element (IR_1) to the tenth element (IR_{10}) is likely very small. In other words, the tenth document is likely to be nearly or just as relevant to the query as the first result.

In contrast, if a query returns only ten results, the drop off in IR score from the first element to the tenth element is likely to be very high.

[0023] The drop off in IR score may be stored in the query breadth database 129 as a ratio, for example, IR_{10} / IR_1 . The breadth of the query is assumed to be highest as the ratio approaches 1, where the relevance of the first and tenth queries are most nearly equal. The breadth is assumed to be lowest as the ratio approaches 0. Another related measure of query breadth is number of results in a result set that have an IR score greater than a percentage of the top IR score, e.g., an IR score greater than about ninety percent (90%) of the top IR score.

[0024] It should be noted that embodiments of the present invention may comprise systems having different architecture than that which is shown in Figure 1. For example, in some systems according to the present invention, server device 104 may comprise multiple physical servers. The system 100 shown in Figure 1 is merely exemplary, and is used to explain the exemplary methods shown in Figures 2 and 3.

[0025] Various methods may be implemented in the environment shown in Figure 1 and other environments according to the present invention. For example in one embodiment, a user 112a enters a search query 114. In response, to receiving the associated query signal 130, a search engine 120 locates documents to return in a result set 134. Before returning the result set 134, the search engine 120 ranks the results. The search engine 120 may sort or rank the results using a popularity measure. In one embodiment, the search engine 120 determines a ranking measure, such as a popularity measure, for a search result in response to the current query. The ranking measure may be or may have been adjusted based at least in part on a query breadth measure of a previously-executed search query associated with the search result. In one embodiment, the adjustment is based on the query breadth measures of a plurality of previously-executed search queries associated with the search result.

[0026] The search engine 120 may also use measures associated with the search query, rather than the number of results returned, to measure the breadth of the search query. For example, in one embodiment, the search engine 120 determines the quantity of search terms in a search query. The fewer the quantity of terms in a search query, the broader the query is likely to be. In another embodiment, the search engine 120 determines how frequently a specific search query is used. The higher the frequency of search query use, the broader the query is likely to be. For example, many users 112a-n are likely to enter a broad query, such as "digital camera." Accordingly, the frequency of use of the query "digital camera" is high.

[0027] The search engine 120 may adjust any popularity measure based on the query breadth measure. The popularity measure that is adjusted may be a query

dependent or query independent measure. A popularity measure that is query independent is a measure of absolute popularity. For example, in one embodiment, a click-through ratio is query dependent since it takes into account the number of times the result is shown to a user 112a as a result of a search query 114. In contrast, the click count is query independent since it is purely based on the number of times an article is clicked.

[0028] Figure 2 is a flowchart illustrating a process for associating a popularity measure with a search result in one embodiment of the present invention. In the embodiment shown, the search engine 120 receives a signal comprising a result, such as an article, and a popularity measure associated with the result at 202. The search engine 120 or other component determines whether or not a popularity measure has been associated with the result by, for example, searching the popularity database (128) at 204.

[0029] If the popularity measure exists for the result, the search engine 120 updates the popularity measure, e.g., the search engine may increment a click count associated with a search result at 206. If the popularity measure does not exist, the search engine 120 creates a popularity measure/result association in the popularity database (128) at 208. The process shown in Figure 2 then ends at 210. The popularity measure can be used in subsequent searches to rank a result returned in response to a search query. The updates to the popularity database 128 may occur offline so that a user is not affected adversely by the storing or updating of the popularity measure.

[0030] Figure 3 is a flowchart illustrating a process for associating a query breadth measure with a search query in one embodiment of the present invention. In the embodiment shown, the query breadth analyzer 126 receives a query signal at 302. In response, the query breadth analyzer determines the breadth of the query at 304. The query breadth analyzer 126 may determine the breadth of the query by various methods. For example, in one embodiment, the breadth of the query is determined as the ratio of the IR score of the tenth result returned by the query divided by the IR score of the first result returned by the query. In another embodiment, the breadth of the query is determined by evaluating the number of results that are returned by the query. A higher number indicates a broader query.

[0031] In one embodiment, the query breadth analyzer 126 receives the information retrieval (IR) score of the first result (IR_1) of the query. The query breadth analyzer 126 also receives the IR score of the nth result (IR_n) of the query. In the embodiment shown in Figure 3, the variable n is set based on previously observed results from queries that are determined in some way to be broad or narrow. The value of n is set to represent the number of results expected from a relatively narrow query. A higher value of n will result in a fewer number of queries being classified as broad, i.e., a fewer number of queries having a slow IR drop rate. In the embodiment shown, n equals 10. The query breadth analyzer 126 next computes the ratio of IR_n to IR_1 (IR_n / IR_1) to arrive at a query breadth measure. To de-weight a result, the computed ratio is subtracted from 1 and multiplied by the popularity measure.

[0032] If the IR score for the query drops off slowly, IR_n will be nearly equal to IR_1 . Accordingly, the ratio of IR_n to IR_1 will approach 1 and the result of

subtracting the ratio from 1 will approach 0. When the popularity measure is multiplied by 1 minus the breadth measure as shown in the embodiment illustrated in Figure 4, the popularity measure will be reduced, i.e., deweighted. Accordingly, results that have unnecessarily high popularity measures as a result of an association with one or more broad queries will be rated less highly in result sets. The query breadth analyzer 126 next associates the measure of breadth with the query, e.g., the query breadth analyzer 126 stores the association in the query breadth database (129) at 306.

[0033] Embodiments of the present invention may utilize query-dependent or query-independent measures for determining the popularity of a result. A query-dependent measure is a measure of popularity that is dependent on the query that returned a result. A query-independent measure provides a measure of absolute popularity for a result, regardless of the context in which the user is presented the result. For example, one embodiment utilizes the click-through ratio for a page to determine the granularity of a page. A very low click-through ratio implies that a result is retrieved in response to many broad queries but is not particularly relevant to users' questions. Accordingly, when the ratio of clicks to impressions is very low, one such embodiment deweights the popularity measure based on the impressions. In another embodiment, the click count is utilized as the measure of popularity. The click count simply measures the number of times an article is clicked without reference to the number of times the article is shown to the user 112a.

[0034] An embodiment of the present invention may utilize a variety of measures of query breadth that are not directly related to the actual results. Such

measures imply the breadth of the query. For example, in one embodiment, a simple measure of breadth, the number of terms used in the search query, provides the measure of query breadth. In general, the fewer the number of terms in the query, the broader the query is. In one such embodiment, the search engine 120 calculates the ratio of the length of the query to the average length of search queries. The lower the ratio, the broader the query.

[0035] In another embodiment of the present invention, the search engine 120 determines the number of times that the exact or almost exact search query has been issued by users 112a-n. The higher the frequency, the more likely the query is to be a broad query. The length of the search query and the frequency of submission of the search query imply that the queries are broad. However, a query may be short or frequent and still relatively narrow. For example, many users may issue a query for a unique surname that has been present in the media. These queries may result in a small number of documents in the result set 134. Therefore, in various embodiments of the present invention, multiple measures of query breadth are combined to determine the breadth of the query and adjust the popularity measure for the search query.

[0036] Figure 4 is a flowchart illustrating a method for adjusting the popularity measure of a result in one embodiment of the present invention. In the embodiment shown, the query breadth analyzer 126 retrieves a record from a log file or other data store that comprises a search query and an identifier of a result that was clicked at 402. The query breadth analyzer 126 then determines a popularity measure for the pair at 404. For example, for a click count, the query breadth analyzer 126

determines how many occurrences of the query/clicked result pair are in the log. The number of occurrences corresponds to the number of times the result was clicked after being retrieved in response to the query. The total number of clicks for the result is a popularity measure corresponding to the search query.

[0037] The query breadth analyzer 126 next determines the breadth of the query at 406. The query breadth analyzer 126 uses the breadth of the query to de-weight the popularity measure corresponding to the search query at 408. The query breadth analyzer 126 stores the de-weighted popularity measure in a data store at 410. The query breadth analyzer 126 or other component may repeat the exemplary process shown in Figure 4 numerous times. For example, the process shown may be repeated for each pair of query / clicked result found in the data store and/or for each popularity measure.

[0038] For example, in one embodiment, a user 112a enters a search query 114 comprising the terms “digital camera.” The client 102a converts the textual search query 114 into a query signal 130 and transmits the query signal 130 over the network 106 to the server device 104. The server device 104 executes the search engine 120 and passes the query signal 130 to the search engine 120. The search engine 120 locates several million documents satisfying the query, “digital camera,” which indicates that the query is a broad query. The query breadth analyzer 126 stores the query, “digital camera,” along with the count of the number of results returned in response to the query in the query breadth database 129. Since the query is broad, many users are likely to enter the same query. The number of times the same query is executed may also be stored in the query breadth database 129.

[0039] In an attempt to narrow the result set, the user 112a subsequently enters a search query 114 comprising the terms “canon powershot digital camera.” The client 102a converts the textual search query 114 into a query signal 130 and transmits the query signal 130 over the network 106 to the server device 104. The server device 104 executes the search engine 120 and passes the query signal 130 to the search engine 120. The search engine 120 locates several hundred thousand documents satisfying the query. The query breadth analyzer 126 stores the query and the count of the results returned in response to the query in the query breadth database 129. Since the subsequent query returns a smaller number of results, the breadth of the query as measured by the number of documents returned for the query “canon powershot digital camera” is smaller than the breadth of the query “digital camera.”

[0040] In, for example, a nightly batch process executed after the two queries described above have been submitted, the ranking processor 124 adjusts the popularity or other ranking measure of a result associated with the previously-executed search queries based on the number of results returned in response to each of the two queries. Although, in the example described, the adjustment is based on only two queries, often, the adjustment is based on a large number of queries.

[0041] Since many of the documents resulting from the broadest exemplary query, “digital camera,” will have been shown many times to many users, the documents are likely to have high popularity measures. However, the popularity measure is artificially high; the score would not be as high if not for the execution of the broad queries. Accordingly, the query breadth analyzer 126 deweights the popularity measures for these documents.

[0042] The adjustment may be based solely on the query breadth measure or may be based only in part on the query breadth measure. For example, the ranking processor may also utilize the age of the popularity data in making the adjustment. Once the popularity measure of the results have been adjusted.

[0043] Subsequently, a user enters another search query, e.g., "canon powershot digital camera Atlanta." The search engine 122 executes the search query and retrieves search results. The ranking processor 124 uses the deweighted popularity measures to sort the results, and the search engine 120 transmits the sorted result set 134 to the client 102a

[0044] The foregoing description of embodiments of the invention has been presented only for the purpose of illustration and description and is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Numerous modifications and adaptations thereof will be apparent to those skilled in the art without departing from the spirit and scope of the present invention.